

# How-To: Load Data into the CDP

- Prerequisites
  - Tools
  - Credentials
- Upload files to S3
  - Login to the SFTP server
  - Upload files to dropbox
  - Copy files into the data lake
- Cleanse the raw data
  - Login to the SSH terminal
  - Run the cleanse command
- Load the cleansed data
  - Run the load command

## Prerequisites

### Tools

1. SFTP client
  - a. mac - <https://cyberduck.io/sftp/>
  - b. windows - <https://winscp.net/eng/index.php>
2. SSH client
  - a. mac - Applications Utilities Terminal
  - b. windows - <https://www.putty.org/>

### Credentials

1. Production SFTP Credentials
2. Production AWS Console credentials
3. SSH access to the production terminal server

[Click Here to Create HELP Request for Access](#)

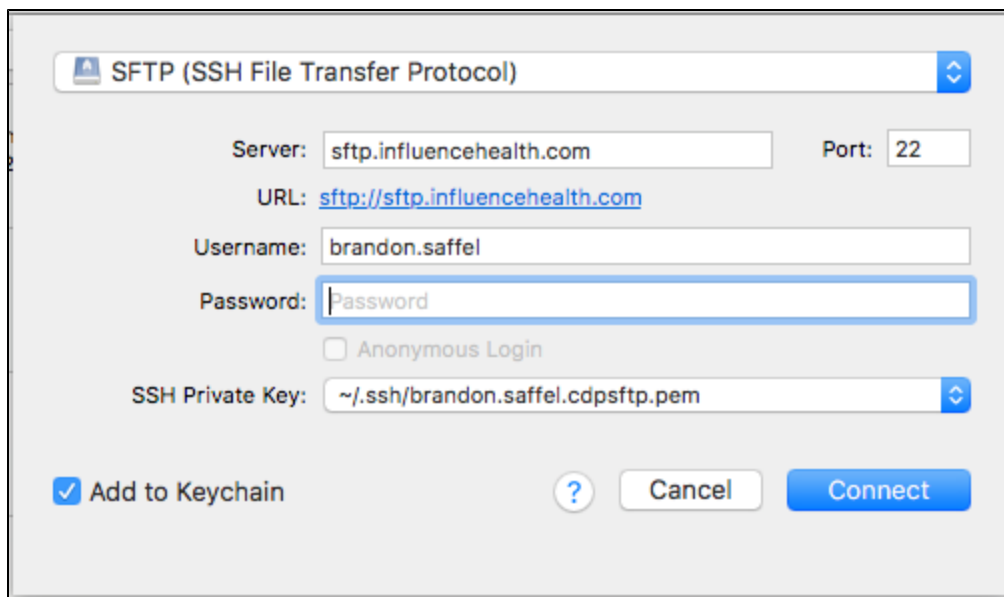
## Upload files to S3

### Login to the SFTP server

SFTP host:

`sftp.influencehealth.com`

Select the correct auth key provided with your access request.



SFTP (SSH File Transfer Protocol)

Server:  Port:

URL: <sftp://sftp.influencehealth.com>

Username:

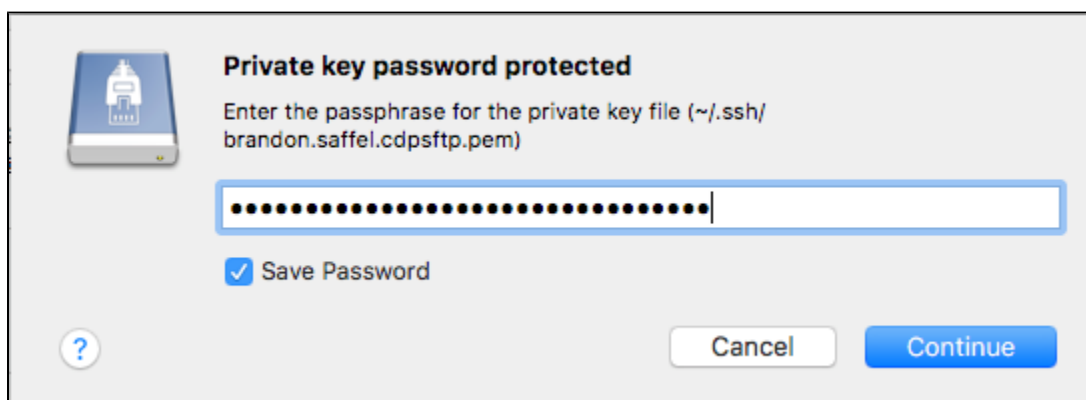
Password:


☐ Anonymous Login

SSH Private Key:

☒ Add to Keychain ?

You will also need to provide a passphrase for the key



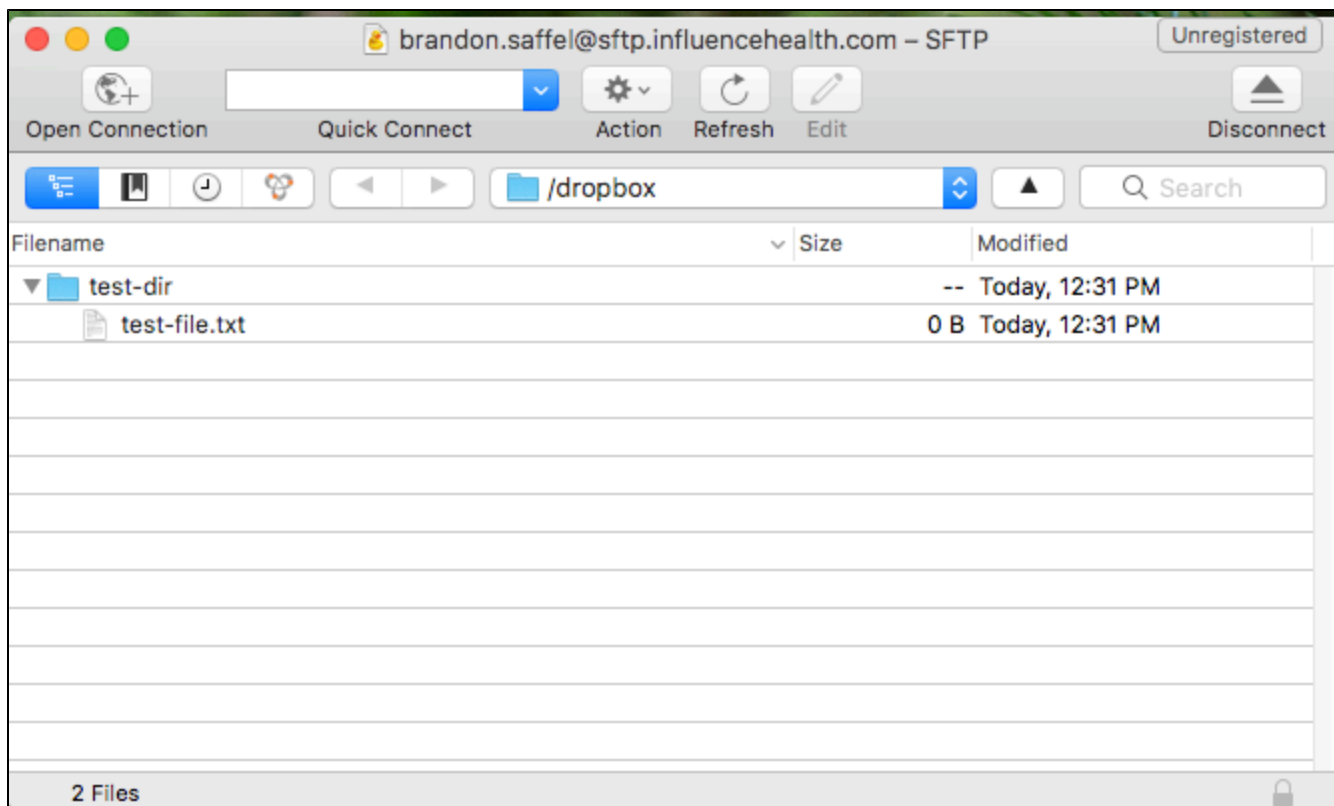
 **Private key password protected**

Enter the passphrase for the private key file (~/ssh/brandon.saffel.cdpsftp.pem)

☒ Save Password

?

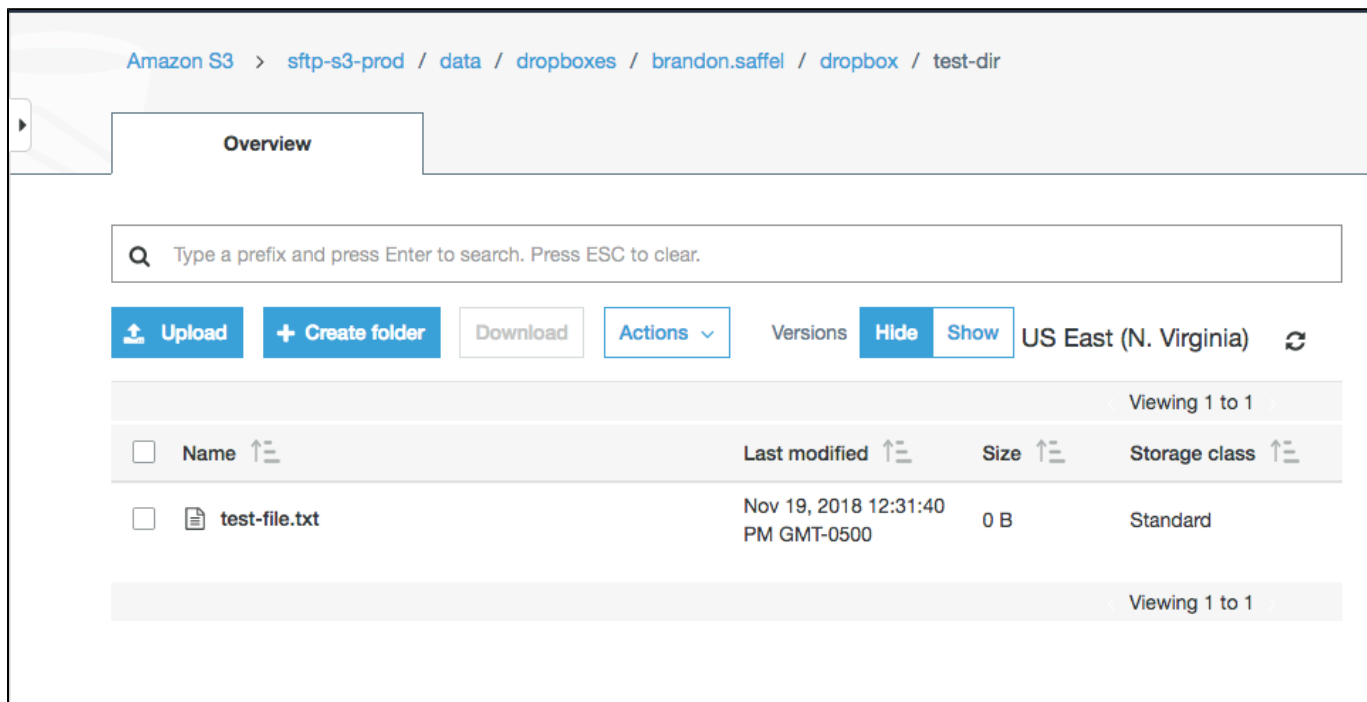
Upload files to dropbox



A step-by-step walkthrough of this can be found at [How-To: Upload Data to S3 via the SFTP](#)

## Copy files into the data lake

1. Login to AWS console and navigate to the correct sftp-s3-prod bucket



Files are immediately uploaded to the dropbox/ directory under your user account. They are then unzipped and copied into the received/ directory.

The original uploaded files are expired/removed each day within the user dropbox. Files within the received/ directory are not

automatically expired and should be cleaned once they have been copied to the data lake

You should be able to see the files that were uploaded in the `data/received` directory after a few seconds to a couple minutes depending on how large the files are. Experian uploads and other multi-gigabyte zip files can take as long as a couple hours.

Copy the files that have been uploaded from the S3 sftp bucket into the correct location in the data lake. Identifying the correct location can be done by following the [Data Lake Naming Conventions for S3 Buckets](#)

Note: This should also preserve the folder structure that was created when files were uploaded, so you can perform a single copy operation for multiple batches at once. In this example 2 batches for tanner are being copied into the correct directory structure at the same time

## Cleanse the raw data

### Login to the SSH terminal

(Temporary) sudo su to root and create (or attach to) a tmux session

More information can be found at [How-To: Run a Cleanse Job](#)

### Run the cleanse command

### Load the cleansed data

### Run the load command